AFRL-IF-RS-TR-2005-412
Final Technical Report
January 2006

STEPS TOWARD THE ALIGNMENT OF COMPLEMENTARY LEXICAL RESOURCES AND KNOWLEDGE DATABASES

International Computer Science Institute

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. Q572

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2005-412 has been reviewed and is approved for publication

APPROVED: /s/

RAYMOND A. LIUZZI
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	JANUARY 2006	Final Nov 03 – Sep 05

4. TITLE AND SUBTITLE
STEPS TOWARD THE ALIGNMENT OF COMPLEMENTARY LEXICAL RESOURCES AND KNOWLEDGE DATABASES

5. FUNDING NUMBERS
C - FA8750-04-2-0026
PE - 62301E
PR - COGV
TA - 00
WU - 02

6. AUTHOR(S)
Charles J. Fillmore and
Collin F. Baker

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
International Computer Science Institute
1947 Center Street
Suite 600
Berkeley California 94704

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFED
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2005-412

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Raymond A. Liuzzi/IFED/(315) 330-3577/ Raymond.Liuzzi@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Since 1997, the FrameNet project at the International Computer Science Institute has been developing a uniquely detailed lexicon of English based on Frame Semantics and “manually” annotated examples from a balanced corpus, and has distributed copies of the lexicon and the annotations to a wide variety of researchers in natural language processing. This contract funded a meeting to evaluate the status of the project and suggest ways in which it could be more useful to the government and other clients. Based on the suggestions of the evaluators, the FrameNet team used the opportunity to increase the coverage of the lexicon, to rapidly add new frame relations and semantic types, to develop software and techniques for full-text annotation, to improve the public website, documentation, data consistency and data distribution system, and to plan for more a more automated workflow and closer connections to WordNet and other knowledge resources, depending on future funding and collaboration with related projects around the world.

14. SUBJECT TERMS
Lexical Resource, Lexical Semantics, Framenet, Wordnet, Cyc, Frame Semantics, Natural Language Processing, NLP, Semantic Role, Semantic Frame, Automatic Semantic Role Labeling, ASRL, Word Sense Disambiguation, WSD, Question Answering, QA, Information Extraction, IE, Argument Structure, Thematic Role, Theta Role, Linking Theory, Ontology, Semantic Web

15. NUMBER OF PAGES
19

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background on the FrameNet Project

It is widely recognized that, while great progress has been made on certain aspects of natural language processing (NLP) in recent years, currently available lexical resources do not provide semantic representations of word senses that are deep and detailed enough to enable the crucial connection from unrestricted text to reasoning and inference systems.

The FrameNet project[1] (hereafter FN), which began in 1997, is providing a way of (a) discovering and recording useful information about contemporary English, in a theoretical framework that takes into account structured information about event types, institutions and artifacts and (b) discovering and displaying the ways in which such information is represented in discourse through individual words and the grammatical ways in which the words fit the phrases and sentences around them. Specifically, FN is building a dictionary of English that is both human- and machine-readable, supported by annotated examples from a corpus, and based on the theory of Frame Semantics.

The intellectual background to the principles and practices of FrameNet is the theory of Frame Semantics (Fillmore, 1968, Fillmore, 1976, Fillmore, 1982, Petruck, 1996) and its assumption that word meanings are best understood in reference to schematic characterizations of situation types, which we refer to as a **semantic frames**, together with a description of the participants, phases and props associated with such situations, which we call **frame elements** (FEs).

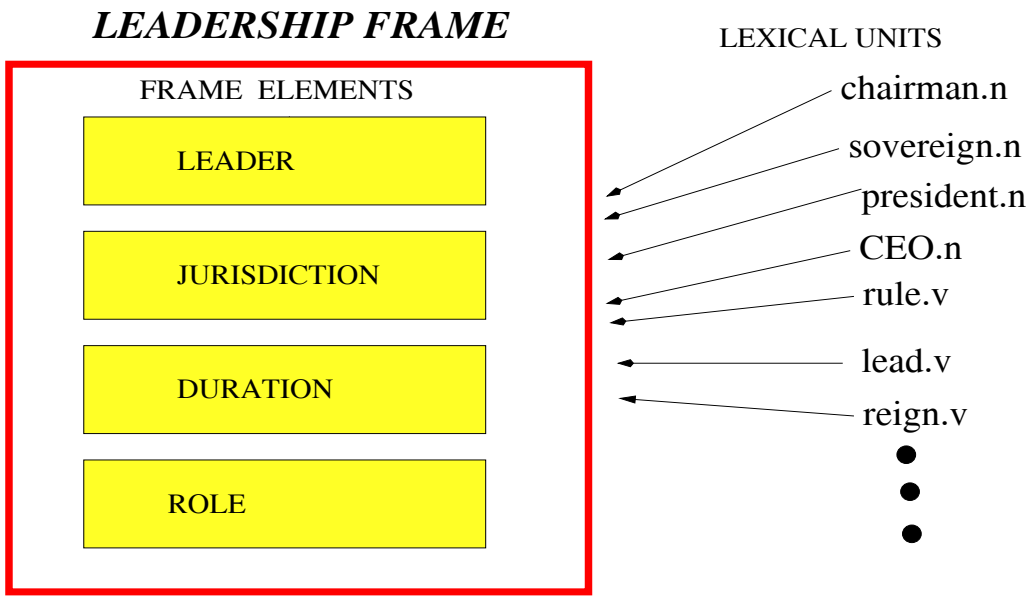

Figure 1 The Leadership frame and some of its FEs and LUs

For example, the Leadership frame (Fig. 1), has to do with a person or agency (LEADER) in a particular ROLE exercising control over some domain (the JURISDICTION). Defining this frame allows us to state just once the largely shared semantics of nouns referring to a title or position (e.g. *king, president, archbishop*) and verbs describing the action of leadership (e.g. *lead, rule,*

1 Fontenelle, 2003, `http://www.framenet.icsi.berkeley.edu`

reign, govern) while individual words have definitions which further differentiate them. (FrameNet is the only computational resource which relates nouns, verbs, and adjectives to the same semantic structures.) Note that ambiguous words (e.g. *ruler* in *ruler of a country* vs. *measure with a ruler*) have to be described in different frames in each of their senses; the FN lexicon treats these separately, as distinct **lexical units** (**LUs**).

The core frame elements of the Leadership frame are the ROLE (kingship, presidency, etc.), the LEADER (i.e., the occupant of the role), the JURISDICTION under the leader's control (nation, bureau, company, etc.), and the DURATION of the control; additional FEs, such as MANNER, MEANS, and DEGREE, may also be significant in sentences that express instances of the Leadership frame. The basic idea fits familiar slot-filler approaches to argument structure: a word evokes a frame; a frame makes available a structure of frame elements with information about their interrelations; particular constellations of frame elements are known to be expressed grammatically in particular ways; and knowing what kinds of semantic material fill the appropriate slots for the frame elements makes it possible to integrate the parts of the conceptual structure built around the frame.

The project has enjoyed seven years of support from the National Science Foundation: first under grant IRI-9618838, March 1997 - February 2000, "Tools for lexicon-building"; then under grant ITR/HCI-0086132, September 2000 - August 2003, entitled "FrameNet++: An On-Line Lexical Semantic Resource and its Application to Speech and Language Technology"; with a smaller but much appreciated supplement in 2004. Since its beginning, the project has been administered through the International Computer Science Institute, in Berkeley, California, an independent organization conducting research in a variety of computer science-related fields.[2]

Instead of relying on standard dictionaries or the intuitions of speakers, FrameNet lexical descriptions are based on actual attested uses of the words under study. For most of the life of the project, FN has depended on the British National Corpus,[3] but has recently been able to add newswire corpus data from the Linguistic Data Consortium.[4]

The BNC and the forthcoming American National Corpus[5] are the results of serious efforts to build balanced corpora. One reason for depending on evidence from very large balanced corpora is that through them we stand a chance of discovering all of the significant combinatory properties of individual lexical items. Other widely available large text corpora like newswire, transcribed casual speech, medical abstracts, and instruction manuals do not provide such an assurance. Once we are equipped with knowledge of the basic valence or combinatorial patterns into which particular words fit, it should in principle become possible to acquire relative frequency information for lexical units (word-sense pairs), frames, frame elements, valence patterns, lexical collocations, etc., in specific corpora of interest, and to show the ways in which these may differ across spoken versus written language, academic versus casual writing, technical versus non-technical writing, etc.

1.2 Purposes of the current agreement

Simultaneous with the FN effort there have been numerous other research efforts directed at providing the research community with reliable resources for NLP including lexicons, ontologies and treebanks. Among these, several have attained a place of prominence in the research of scholars everywhere: the WordNet lexicon, which shows a wide range of word-to-word relations,

2 `http://www.icsi.berkeley.edu`
3 `http://www.hcu.ox.ac.uk/BNC`
4 `http://www.ldc.upenn.edu`
5 `http://www.americannationalcorpus.org`

such as synonymy, antonymy and hyponymy for more than 150,000 word senses,[6] the Cyc effort to build a large ontology and knowledge base for common-sense reasoning,[7] and the PropBank annotated treebank, which links syntactic form to argument structure.[8]

Contract FA8750-04-2-0026 was awarded to the International Computer Science Institute (ICSI) Dec. 31, 2003 for the purpose of evaluating the feasibility of aligning, integrating, or otherwise making a connection between the FrameNet lexical database and other resources, particularly WordNet and Cyc. The three major steps in this process, as outlined in the "Task" section of the Statement of Work were:

(1) to prepare and conduct a workshop, to be held at ICSI, bringing together experts in computational linguistics and NLP, to evaluate the work of FrameNet,

(2) based on the advice contained in this report, to conduct a feasibility study on the integration of FrameNet, WordNet and Cyc, and

(3) to prepare and release a new version of the FrameNet data in RDF/DAML/OWL format.

1.3 FN full-text annotation

In recent years, FN has begun annotating continuous texts from several sources, as a kind of benchmark for deep semantic annotation. Although the project initially concentrated on lexicon-building, there was always an expectation that the frames and frame elements being developed would be used to "densely" annotate continuous texts for various NLP purposes, as discussed in Lowe et al. 1997, Fillmore and Baker 2001, etc. Since 2000, the database has contained a means of representing hierarchical document structure at the levels of corpus, document, paragraph and sentence, although these were little-used so long as the sentences were extracted individually from the BNC solely as lexicographic examples.

It should be made clear that during the period of this contract (12/31/2003–5/31/2005, which includes an extension), the members of the FrameNet team were also working on two partially overlapping tasks. The first was a subcontract on NSF grant IIS-0325646 (Dan Jurafsky, P.I.) "Domain-Independent Semantic Interpretation" (9/1/2003–8/31/2005); for providing full-text FrameNet-style annotation of texts that had also been annotated in the PropBank project; it is hoped that this will be the first step toward establishing a correspondence between the two annotation projects (cf. Ellsworth et al., 2004, Palmer et al., 2005). Work on this grant represented the first time that FN had annotated continuous text, although the theory behind it had been developed and discussed for many years.

The second task was a subcontract under an ARDA award to Sanda Harabagiu (University of Texas at Dallas) and Srinivas Narayanan (ICSI) to work on question answering (9/27/2004–3/26/2007), as part of the AQUAINT project. FN staff are now annotating AQUAINT texts, and a set of passages, with questions and answers about them with FN frames and FEs, testing whether this will improve the results of the QA system. Work on both of these tasks, like the work on the task described herein, has enlarged and improved the FrameNet lexical database; lexical units added in the context of one task are then available for use in other tasks. Nevertheless, the following discussion will be confined as much as possible to work done under DARPA contract.

---

6 `http://wordnet.princeton.edu`

7 `http://www.cyc.com`

8 `http://www.cis.upenn.edu/~ace`, (Palmer et al., 2005)

2 The FrameNet Lexicon and Annotation Process

In contrast with the work of standard commercial lexicography, which proceeds through the dictionary's word list alphabetically, recording all meanings of each word, the lexicographic work at FrameNet has been conducted on a frame-by-frame basis; the basic steps are:

1. Select a semantic area and outline the frames involved, through a combination of the analysts' intuitive knowledge as native speakers and careful examination of corpus data.
2. Define the frames and their frame elements (FEs) and prepare a list of lexical units (LUs), with a brief definition of each. It is usually possible to define new frames in relation to the existing taxonomy of frames (Sec. 2.1).
3. For each LU, find the principal syntactic patterns and common collocations and extract examples of each from the corpus.
4. Annotate the examples to provide evidence for all the syntactic realizations of each FE.
5. Run report software to view the **valence** patterns of each LU along with the LU definition, which together constitute the **lexical entry** for the LU.

2.1 The Frame Taxonomy

Neither WordNet nor FrameNet began as an effort to create the ontology, but they have shared the fate of seeing other researchers treat their products as ontologies. In FrameNet's case there were projects that simply used frame descriptions as a basis for classification of events (e.g. Porzel et al., 2003). But FrameNet researchers had also seen the need to formalize frame-to-frame relations, in order to represent the idea that that some frames are parts or subtypes or analogues of other frames; in other words, FN has had to define semantic frames that would facilitate generalizations over groups of LUs (and FEs).

As part of that process, a desire to capture higher-level generalizations has led to the definition of more general frames, with several types of relations being established between frames.[9] Tools for editing these relations and a web-based viewer have been developed. The most important of these relations are discussed below.

2.1.1 Inheritance

We define frame **inheritance** as an IS-A relation between a parent frame and a child frame which includes full inheritance of FEs and their semantic types (discussed in Sections 3.3 and 3.5). This means that if the parent frame has a semantic type, the child frame must have the same semantic type or a subtype (elaboration) of it. Also, for each FE in the parent frame there must be an FE in the child frame of the same semantic type or a subtype thereof. The FEs of the child may or may not have the same names as the FEs of the parent, and there may be additional FEs in the child for which there is no corresponding FE in the parent. Furthermore, if the parent frame has subframes (basically subevents, discussed below), its subframe structure is also inherited (and possibly elaborated) by the child frame. This is complete, monotonic inheritance.

9 This has caused some users of the FN data to regard it as an ontology; as with WordNet, this was not the original motive in creating the resource.

2.1.2 Using

Because we found a number of relations among frames which do not quite fit the criteria for full inheritance, we have defined a second type of relation, called **using**, similar to inheritance, but which does not require full mapping of FEs from parent to child, or complete inheritance of the parents' subframe structure. Like inheritance, there can be multiple using relations, so that a child may inherit and/or use multiple parents.

Defining such inheritance relations creates a lattice of frames, a directed acyclic graph. Working out the full details of multiple inheritances through several levels of the lattice can be complex and time-consuming, and we have not created all the links which we would like, but progress is being made.

2.1.3 Subframes

Subframes are used for representing subevents; frames that represent complex processes have subframes representing their subparts[19]. To take a simple example, the Motion_scenario frame has three subframes, Departing, Motion, and Arriving. In this case, the subframes are temporally ordered, bu tin general subframes need not be completely ordered with respect to each other.

For example, the Commercial_transaction frame has two subframes Commerce_goods-transfer and Commerce_money-transfer, but these are not ordered with respect to each other. In some commercial transactions, you pay in advance, in others, only after receiving the goods or services. Furthermore, each transfer can independently occur in time-distributed variants, as with time-payments (for money transfer) and distributed product delivery (for goods transfer), such as quarterly payments for a newspaper delivered daily.

2.1.4 Implementation of Frame Relations in the Database

The conceptual relations described above are implemented as a relational database, using MySQL. So far as possible, the tables of the database and the links between them directly mirror the conceptual structure. For example, there is a table for frames and another for FEs, with a one-to-many relation between them. The lemma table is linked to the frame table via the lexical unit

[19] Note that **subtypes** of frames are represented by inheritance and using relations, which are quite distinct from the **Subframe** relation.

table, each entry of which has a pointer to a lemma and a pointer to a frame; this is a many-to-many relation—frames typically include many lemmas and the same lemma can appear in several frames, representing polysemy or homonymy. The higher-order relations are handled similarly. For a full discussion of the database structure, see Baker et al., 2003.

3 Results and Discussion

3.1 The FN Evaluation Meeting, December, 2003

The meeting described in the Agreement took place Dec. 8-9, 2003, attended by the following:

DARPA/IPTO:	Ron Brachman, David Gunning, Murray Burke
FrameNet:	Charles Fillmore, Collin Baker, Josef Ruppenhofer, Michael Ellsworth
WordNet:	George Miller, Christiane Fellbaum
Cyc:	Doug Lenat, Michael Witbrock
Xerox PARC:	David Israel, Danny Bobrow
Stanford	Chris Manning
U MD/MIACS	Philip Resnik
	John Sowa

The discussion from the meeting continued on line and culminated in a document "Report from the Berkeley FrameNet Review Meeting", edited by Michael Witbrock and submitted to Ron Brachman, Director of DARPA/IPTO in April, 2004. The report identified three major areas of concern and recommended that the FrameNet project concentrate on making progress in these regards:

1. Coverage of linguistic phenomena
2. Applicability to current research problems
    (a) Suitability
    (b) Ease of Use
3. Integration with other lexical, syntactic, and semantic resources

There was some disagreement among the members of the committee as to how the coverage of FrameNet could be increased more rapidly (based on the idea that comprehensive NLP systems require information on roughly 60,000 words) without sacrificing the quality of the annotation. There was also concern over whether the selection of clear, simple examples required for lexicographic work would provide suitable training data for machine learning of systems for automatic frame recognition and FE labeling.

Another concern was whether the choice of frames and LUs, and the texts used to exemplify them were appropriate for DARPA funded research. Finally, there was a concern that the FN data in its current form was relatively difficult to understand and make use of.

3.2 Summary of the current coverage of the FN lexicon

Frames	752	
Lexical Units		
	Fully Annotated	**Total**
Verbs	2394	3970
Nouns	2421	4027
Adjectives	1129	1714
Other POS	19	155
Total	5963	9866

Table 1: Current size of the FN lexicon

Table 1 shows some current statistics for the FN lexical database. As mentioned above, FN treats words of all parts of speech that express the same concept as being in the same frame; it is obvious from the numbers above that most of the work thus far has been concentrated on nouns, verbs and adjectives, but there is no theoretical reason why more cannot be done on the frame semantics of adverbs, prepositions, etc. As can be noted from the table, roughly 60% of the LUs have been "finished" lexicographically, i.e. each has roughly 20 fully annotated example sentences in the lexical database. Many of the others have been annotated in the full-text annotation, but there may be only one or two examples, and these are not included in the lexical database. There are slightly more than eight FEs per frame on average, of which roughly half are very general types of FE such as place and time, and the other half are more or less unique to the frames they occur in, such as avenger and punishment in the Revenge frame.

3.3 Aligning FN with WN

It might seem that the frame taxonomy would map directly onto existing ontologies. In the early days of the project, the FN team had hoped to be able to populate a given frame simply by bringing in all the words in a WordNet synset. In fact, the first draft of the very first proposal, back in 1996, expressed the intention to build FrameNet on top of WordNet, and to make it available as part of the distribution of WordNet. But this has proved illusory; the division of LUs in FN on the basis of which groups of FEs they relate to syntactically produces different groupings from WordNet's sets of synonymous words **synsets**. FrameNet staff members do however, use WordNet synsets as a source of suggestions as to words which might participate in a given frame, along with other thesauruses both on paper and on-line.

Some of the most interesting work in this direction has been done by the members of the SALSA project at Saarbrücken, with whom we have a close working association, aided by a grant from the Alexander von Humboldt Foundation. In a 2005 paper entitled "A WordNet Detour to FrameNet", Aljoscha Burchardt, Katrin Erk, and Anette Frank demonstrate that the problem of the sparsity of FN data can be partially overcome by looking up words missing from FN in WordNet and finding the closest synset that is comparable to the set of LUs in a FN frame. There is also a web site based on this sort of mapping, where a user can type in a word then choose the WordNet synset which most closely represents the meaning she has in mind, and be presented with the closest matching FN frame and LU. (http://www.coli.uni-saarland.de/~albu /cgi-bin/FN-Detour.cgi, accessed Nov. 14, 2005 at 01:42 GMT)

In connection with the development of the frame taxonomy, FrameNet has created a set of 57 semantic types, which can be applied to FEs, frames, or LUs, and used these to specify semantic types for many FEs; multiple inheritance and multiple semantic types on a single entity are allowed. These semantic types include types that point to selected noun nodes in WordNet; the hope is that FEs marked with these semantic types will be realized by NPs headed by nouns in synsets which are hyponyms of these nodes in WordNet[20]. Similar methods, based on treating the WordNet hierarchy as a tree of sorts (mainly on nouns), have been used by other researchers such as Resnik 1993 and Green 2004, and the concept is related to the method used by Mohit and Narayanan 2003 to enlarge the list of nouns for IE patterns by following head nouns from FN examples up the WordNet hierarchy.

Other researchers have also produced resources combining WordNet and FN, or used such a combination in their work. Rada Mihalcea and her student Lei Shi have combined FrameNet and WordNet for automatic semantic role labeling (ASRL), also known as "semantic parsing" (Shi and Mihalcea, 2004, 2005), using WordNet nodes as semantic types for FEs. Alessandro Moschitti and collaborators have worked on ASRL using both FN and PropBank (Moschitti and Basili, 2005), and also combined these resources for knowledge discovery (Giuglea and Moschitti, 2004).

3.4 Aligning FN with CYC or other ontologies

A number of users of the FN data, along with several members of the evaluation committee, have expressed the desire for a link between FN annotations and some formal semantic ontology, such as Cyc. The project team has explored certain portions of the publicly-available portion of the Cyc database, for ideas on alignment possibilities, but without much success. Prior to that, there was a period of discussion with Adam Pease and others on the possibility of linking the FN vocabulary with the SUMO ontology. It quickly became clear that there is a difference between grouping objects and concepts into a taxonomy for AI purposes and grouping **linguistic** entities and expressing their relations to each other.

However, there have been two recent contributions to FN from visiting post-doctoral researchers which have helped the project move toward a more formal system. The first of these is the work of Francisco Valverde (Valverde-Albacete, 2005), who studied the frame taxonomy in the light of formal concept analysis and lattice theory, by treating the frames as nodes in a lattice with FEs as their attributes. He extracted the relevant parts of the FN database and loaded them into standard lattice visualization software. His analysis noted (1) that the FN lattice tends to be very "flat", meaning that there are very few higher-level frames which then have a large number of frames as their children, and (2) that our treatment of FEs that represent portions of the Path in frames in the Motion domain is inconsistent with the treatment of other FEs[21] These results suggest that (1) there may be a need to create more intermediate-level frames and (2) there may be a need to define a new sort of entity, an "proto-role" or "proto-FE", which can then be realized by multiple actual FEs.

The second contribution is the work of Jan Scheffczyk, who has written a set of rules in first-order predicate logic which express a number of principles of frame semantics in general and

11 Actually, a semantic type can also be satisfied when a neutral noun is **modified** by an adjective related to the type. Thus, for contexts calling for negatively evaluated things, we find, in addition to *disaster, tragedy, accident, infection,* etc., also phrases like *disastrous situation, unhealthy condition,* and *bad healt*".

21 The path FE is unusual in many respects; for example, it is the only FE that can legitimately occur several times in the same clause, as in *The ball **flew** [over third base], [through the open window] and [into the pot of soup on the stove]*.

FrameNet in particular, such as “Every frame must have at least one FE” and “In an inheritance relation, every core FE of the parent frame must be bound to a child FE in the child frame via an FE-FE relation.” He is then algorithmically testing the consistency of the FN lexical database against these rules and reporting any violations. It is expected that this process will enable the next release of the FN data to be far more consistent with the principles underlying the project. Finally, the release of the FN data in RDF/OWL format (discussed in Sec. 3.6), enables the data to be loaded into standard tools for ontology-building.

3.5 Enhanced Frame-to-Frame relations and Semantic types

A major effort has been made as part of this grant to increase the number and improve the accuracy of frame-to-frame relations, with their accompanying FE-to-FE relations; between Release 1.1 (Jan, 2004) and Release 1.2 (June, 2005) the number of frame-frame relations increased by one third, from 683 to 914, (with a similar increase in the number of FE-FE relations).

One of the benefits of creating a graph with higher connectivity is that semantic typing of FEs can be made more complete; over the period of this contract, the number of FEs bearing semantic types increased from 49 to 1,528[22] . This was largely accomplished by defining FE-FE relations and then propagating semantic types from parent to child down the hierarchy. Increased semantic typing on FEs should make it easier for NLP systems to recognize which constituents fill which FE roles (e.g. *The train arrived in the village* vs. *The train arrived in the morning*) and even which frame is being evoked by polysemous words (e.g. *He tied the game* vs. *He tied the knot*).

3.6 FrameNet Data Releases

Since the primary product of the FrameNet project is the descriptions of the frames and FEs and the sentences annotated on the basis of these descriptions, a major activity has been collecting this data, putting it into a format suitable for use by others, and releasing copies to interested users. There have been four releases[23] of the FrameNet data since the beginning of FrameNet II, as shown in the table above; with each release ideas about what sorts of data should be released and what formats are most useful has been refined, based on feedback from the growing user community. During the course of this contract, a new, more automated system for data distribution was also written and put into use.

Release No	LUs	Frames	FEs	Fr. Rels.	Anno. Sets
0.7	4,281	182	-	-	84,851
1.0	6,904	384	3,059	-	127,293
1.1	7,859	487	3,936	352	132,968
1.2	8,755	607	4,908	485	133,846

Table 2: The growth of the FN database releases

Notes: Release 1.1: Beginning with release 1.1, the count shown in the right-hand column is not for sentences, but for annotation sets; Some sentences were annotated with more than one

22 Unfortunately, this increase came shortly after the cutoff date for R1.2, but the new types will be included in the next release, 1.3.

23 In addition to the formal releases in XML and HTML, the data in SQL dump format has been given to Prof. Hans C. Boas of U of Texas at Austin (who is interested in building a German FrameNet) and to Dr. Petra Steiner at U of Erfurt (Germany), where she is using it for studies on semantic clustering.

annotation set, and, beginning with this release, we had a good representation for this situation, reflected in changes in both the database and the XML format.

Release 1.2: Although the total number of annotated sentences in R1.2 did not increase significantly from R1.1, a major effort was made to eliminate errors and improve the consistency of annotation. This included both deleting some questionable sentences and adding new annotation to existing sentences. In particular, sentences with verbs as targets were reviewed to make sure that every core frame element of the verb is either labeled on a portion of the sentence or marked as "null instantiated"; this will provide data for research on discourse cohesion and anaphora resolution that is otherwise very difficult to obtain–information about what is **missing** from a sentence, often with an idea of its semantic type.

Occasionally it is necessary to split or join existing frames, and move annotated sentences from frame to frame in this process. Release 1.2 contains version difference files to help current users track these changes and bring their applications into alignment with the new frames and LUS.

Beginning with Release 1.2, the frame, FE and frame relation data is being distributed not only in the XML format defined by the FN project, but also in RDF/OWL format. This can be read by standard ontology tools such as Protege, affording a more formally structured presentation of the data, and the means for editing it for purposes such as building an ontology or comparing it with other ontologies. This should be of use to researchers working on the Semantic Web, and to anyone who wants a more formally expressed semantics for FN, although it is certainly less human-readable than even the XML format.

Although many users seem perfectly satisfied with the XML data files and report that they are easy to parse and extract information from, others have requested a more convenient interface, one that would make it easier to extract particular relations from the data. Accordingly, the next data release R1.3 (scheduled for December, 2005), is planned to contain a Java API, which will provide both methods that will be easy to call from within other Java programs and a command-line interface. A beta version of this API is now being tested.

3.7 Users and Uses of the FN data

The FrameNet data have been very widely used by a great variety of researchers. Each of the recent releases (1.0, 1,1, and 1.2) has been downloaded hundreds of times; the users include NLP researchers working on information extraction, machine translation, question answering and automatic semantic role labeling, computational linguists seeking a lexicon with information on subcategorization and selectional preferences, college-level teachers of computational linguistics or lexical semantics, and individuals doing linguistic research. After U.S. universities, the most common institutional affiliations are with universities in Europe, India, and the Far East, followed by research laboratories (both government and industrial). A selection of roughly 150 users' names, affiliations and purposes is posted on the project website (with each user's permission)[24] . In the process of requesting to download the data, users are asked about their areas of interest; the distribution of replies looks like this (multiple replies are permitted):

24 http://framenet.icsi.berkeley.edu, choose "Users" on main menu.

90	Teaching lexical semantics
121	Machine translation
135	Natural language generation
151	Lexicography
189	Question answering
235	Word sense disambiguation
243	Research in lexical semantics
273	Information extraction
289	Semantic parsing
334	Natural language understanding

Table 3: Data users' fields of interest

In addition, there are funded research teams who are creating FrameNets for Spanish[25] and Japanese[26] and a major text annotation project for German[27] which uses FN frames and FEs where available; joint meetings of these projects take place on a regular basis. Efforts are also being made to extend FN to other languages, including Chinese, French, and other Romance languages.

3.8 Rebuilding of the FN website

One of the recommendations of the evaluation committee was that the project should do a better job of explaining its work and its data to the public. A major improvement in this regard, has been the complete overhaul of the project website, `http://framenet.icsi.berkeley.edu`, which is now more understandable for the first time visitor. New features include

- a forum where users can leave questions for the staff
- display of all data, current within 1 month (except for certain grammatical information) including full-text annotation
- a graphical browser for frame-frame and FE-FE relations, allowing users to start from any frame and select the relations to be displayed and the number of links to follow from the starting frame, and
- a revised version of the FN annotation manual, which explains the theory and practice of annotation

3.9 Future plans: Computer-assisted Annotation and Vanguarding

As noted above, a major concern of the evaluation committee was the relatively slow growth of FrameNet in terms of the number of words covered, and part of the reason to seek connections with WordNet (in particular) was to quickly provide "coverage" for a much larger number of words. One obvious approach would be to find ways to automate the FN process itself, both the process of defining frames and extracting example sentences (**vanguarding**) and the process of annotating either the lexicographic examples or the sentences of continuous text.

One promising new idea is represented by a recent grant proposal by Nancy Ide (ANC - Vassar) and Christiane Fellbaum (WordNet - Princeton) with FrameNet (Baker - ICSI) as a subcontractor, which would enhance the American National Corpus[28] by including hand-validated annotation of a 10 million word balanced sub-corpus for syntax, named entities, and semi-automatic annotation of WordNet senses and FrameNet LUs and FEs. Thus, this text would relate WordNet senses and FN LUs in the context of sentences fully annotated in multiple

25 Spanish FrameNet, `http://gemini.UAB.es/SFN`
26 Japanese FrameNet `http://jfn.hc.keio.ac.jp`
27 The SALSA project, `http://www.coli.uni-saarland.de/projects/salsa`
28 The ANC is under construction; a 10 million word subset of the planned 100 million words has been released by the Linguistic Data Consortium, and a 20 million word subset is forthcoming.

frameworks. The FrameNet part of this plan depends heavily on recent work on automatic frame recognition (Erk and Padó, 2005, Erk, 2005), and FN-based automatic semantic role labeling (e.g. Baldewein et al., 2004, Burchardt et al., 2005, also discussed in Sec. 3.3)

Another proposal seeks funding for partially automating the vanguarding process. The current system requires separate stages of interactively searching the corpus for examples of usage and then defining a set of rules for extracting the examples;[29] if a polysemous word is being studied, the work done on one LU (word sense) does not carry over to the other(s). The proposed system would build on the WordSketch system of Adam Kilgarriff (Kilgarriff et al., 2004), which divides clusters of examples of word usage based on the occurrence of statistically significant correlates in key syntactic positions, and presents these clusters to the user so that they can be marked interactively as different word senses. The proposed FN system would use such an interface and tie it to the frame definition process, so that such clusters could be converted into rules for example extraction, including automatic marking of FEs on certain constituents. As the vanguarding is more time-consuming than the annotation, this proposal could considerably speed up the growth of the FN lexicon.

4 Conclusions

The FrameNet project is producing both a uniquely rich semantic lexicon and fully-annotated texts in domains of interest to the government. These are being widely used in the research community, and show promise of improving the performance of real question answering and information extraction systems. The evaluation committee raised a number of legitimate concerns about the progress of FrameNet and its relation to other work in the field; most of these concerns have been addressed in work done under this contract, while others are addressed in proposals awaiting funding. It is to be hoped that continued support will enable the FrameNet team to return to its previous strength and further develop this valuable resource.

29 cf. Fillmore et al., 2003 for details of this process

References

Baker, C. F., Fillmore, C. J., and Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3).

Baldewein, U., Erk, K., Padó, S., and Prescher, D. (2004). Semantic role labelling with similarity-based generalization using em-based clustering. In Mihalcea, R. and Edmonds, P., editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 64–68, Barcelona, Spain. Association for Computational Linguistics.

Burchardt, A., Erk, K., and Frank, A. (2005). A WordNet detour to FrameNet. In *Proceedings of the GLDV 2005 workshop GermaNet II*, Bonn.

Ellsworth, M., Erk, K., Kingsbury, P., and Padó, S. (2004). PropBank, SALSA, and FrameNet: How design determines product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal.

Erk, K. (2005). Frame assignment as word sense disambiguation. In *Proceedings of IWCS 6*, Tilburg.

Erk, K. and Padó, S. (2005). Analysing models for semantic role assignment using confusability. In *Proceedings of HLT/EMNLP-05*, Vancouver, Canada.

Fillmore, C. J. (1968). The case for case. In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*. Holt, Rinehart & Winston, New York.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280:20–32.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL.

Fillmore, C. J., Petruck, M. R. L., Ruppenhofer, J., and Wright, A. (2003). Framenet in action: The case of attaching. *International Journal of Lexicography*, 16(3):297–332.

Fontenelle, T., editor (2003). *International Journal of Lexicography*, volume 28. Oxford University Press. (Special issue devoted to FrameNet.).

Giuglea, A.-M. and Moschitti, A. (2004). Knowledge discovery using framenet, verbnet and propbank. In *Proceedings of the Workshop on Ontology and Knowledge Discovering*, Pisa, Italy. ECML.

Green, R. (2004). *Inducing Semantic Frames from Lexical Resources*. PhD thesis, University of Maryland, College Park.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (July 2004). The sketch engine. In *Proceedings of EURALEX 2004*, Lorient, France.

Lowe, J. B., Baker, C. F., and Fillmore, C. J. (1997). A frame-semantic approach to semantic annotation. In *Tagging Text with Lexical Semantics: Why, What, and How? Proceedings of the Workshop*, pages 18–24. Special Interest Group on the Lexicon, Association for Computational Linguistics.

Mohit, B. and Narayanan, S. (2003). Semantic extraction with wide-coverage lexical resources. In Hearst, M. and Ostendorf, M., editors, *HLT-NAACL 2003: Short Papers*, pages 64–66, Edmonton, Alberta, Canada. Association for Computational Linguistics.

Moschitti, A. and Basili, R. (2005). Verb subcategorization kernels for automatic semantic labeling. In *Proceedings of the ACL05 Workshop on Deep Lexical Acquisition*, Ann Arbor, MI.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *cl*, 31(1):71–106.

Petruck, M. R. L. (1996). Frame semantics. In Verschueren, J., Östman, J.-O., Blommaert, J., and Bulcaen, C., editors, *Handbook of Pragmatics*. John Benjamins.

Porzel, R., Gurevych, I., and C., M. (2003). Ontology-based contextual coherence scoring. In *Proceedings of the 4th SIGDial Workshop on Discourse and Dialogue*, Sapporo, Japan. Association for Computational Linguistics.

Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania.

Shi, L. and Mihalcea, R. (2004). Semantic parsing using framenet and WordNet. In *Proceedings of the Human Language Technology Conference (HLT/NAACL)*, Boston.

Shi, L. and Mihalcea, R. (2005). Putting the pieces together: Combining framenet, VerbNet, and WordNet for robust semantic parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico.

Valverde-Albacete, F. J. (2005). Explaining the structure of framenet with concept lattices. In Ganter, B. and Godin, R., editors, *Proceedings of the Third International Conference on Formal Concept Analysis (ICFCA 2005)*, volume 3403 of *Lecture Notes in Computer Science*, pages 79–94, Lens, France. Springer-Verlag GmbH.

Appendix: FrameNet Glossary

By Charles J. Fillmore and Miriam R. L. Petruck

annotation database - the part of the database that holds the annotated sentences

annotation - the assignment of semantic role tags to syntactic constituents

FN - FrameNet

frame (semantic frame) - a schematic representation of a situation involving various participants, props, and other conceptual roles, each of which is a frame element**frame semantics** – a descriptive framework for characterizing lexical meaning in terms of semantic frames

frame element (FE) - frame-specific defined semantic role that is the basic unit of a frame

inheritance - a frame-to-frame relation in which the child frame elaborates the parent frame; the child frame is said to be a "kind-of" parent frame, e.g. Arriving is a kind-of Motion

full inheritance - each core FE in a parent frame is bound to an FE in the child frame, though not necessarily one called by the same name

monotonic inheritance - inherited characteristics cannot be overridden

multiple inheritance - a child frame (and therefore its FEs) can have more than one parent

lemma - a unit made up of one or more lexemes seen as bearing one or more senses, e.g. *bring up* consists of the lexemes *bring* and *up*

lexeme - a word in a given part of speech instantiated by one or more word-forms, e.g. the lexeme *bring* has the word forms *bring*, *brings*, *bringing*, *brought*

lexical database - the part of the database that holds the frames, frame elements and lexical units

lexical entry – the syntactic realization of the frame elements and the valence patterns of a lexical unit

lexical unit (LU) - a pairing of a lemma and a frame - i.e. a "word" taken in one of its senses, e.g. the verb *tie* in the ATTACHING frame

null instantiation - a missing frame element that would normally be expected in a sentence, e.g. in *She already told him*, the Message is considered to be understood in context, but is not overtly expressed in the sentence, i.e. it is "null instantiated".

POS - **part of speech**

semantic type - a mechanism used to capture semantic facts about individual frames, FEs, and LUs that don't fit into the developing hierarchy of frames in FrameNet

semantic valence – the frame that underlies the meaning of a word, and the number and kinds of entities that participate in the situation instantiating the frame

subframe - a frame-to-frame relation whereby (smaller) component frames comprise parts of a (larger) complex frame

syntactic valence – the number and type of syntactic constituents that are dependent on, or in construction with a word

target - the lemma under consideration, and in respect to which annotation is provided

uses - a frame-to-frame relation like Inheritance, but less strictly defined

valence - the particular kinds of constituents, in terms of semantic roles, grammatical functions, and phrase types, with which a word combines in a grammatical sentence

valence pattern - the set of valence groups realized in one sentence

XML - Extensible Markup Language, widely used as a format for the exchange of data between different computer systems, programs, etc.